

本周周报(12.17-12.23):

解聪

本周工作:

1. 时序用户行为分析

本周实现了时许数据可视化对 `twitter` 数据的初步验证。

数据使用汪飞的 `twitter` 数据，对于每条记录，先提取了如下属性：

``id``, ``userId``, ``username``, ``content``, ``creationTime``, ``replyToPostId``, ``replyToUsername``, ``retweetedFromPostId``, ``retweetedFromUsername``, ``retweetCount``

分别代表：`tweet` ID，用户 ID，`tweet` 内容，发出时间，所回复 ID，所回复的用户名，所转发的 ID，所转发的用户名，被转发的次数。

原始数据还包括来源，地点等等。但是地点数据太稀疏，所以先剔除掉了。

因为处于快速验证方法的阶段，所以很多工作其实是手动完成的，比如主题提取。提取了一天中比较热门的 38 个话题，并就各条记录所对应的话题类别手动分类。

对提取出的记录可视化映射的方式如下：

图标	含义
符点	某个用户的一组 <code>twitter</code>
符点颜色	感情偏好
符点大小	被转发/评论的数量
音符组	话题
符尾颜色	话题类别（政治，经济等）
音符上方文字	关键词，来源
横向布局	时间
纵向布局	无（需改进）

1. 微博的话题

微博的主题抽取由人工完成，其次将每条被回复或被转发比较多的微博当作一个话题，对筛选出的近 700 条转发比较多的微博相关领域进行了手动分类，比如政治，经济等。

每个话题以一组音符的形式展现，不同题材的话题其符尾颜色不同。如图 1 中被选中的显示的是有关计算机病毒的一个话题。每个话题上方会出现与之相应的关键词(关键词提取算法有待改进)，以及话题发起者。

通过音符出现的位置可以判断该话题出现的时间点以及延续的时间长短。

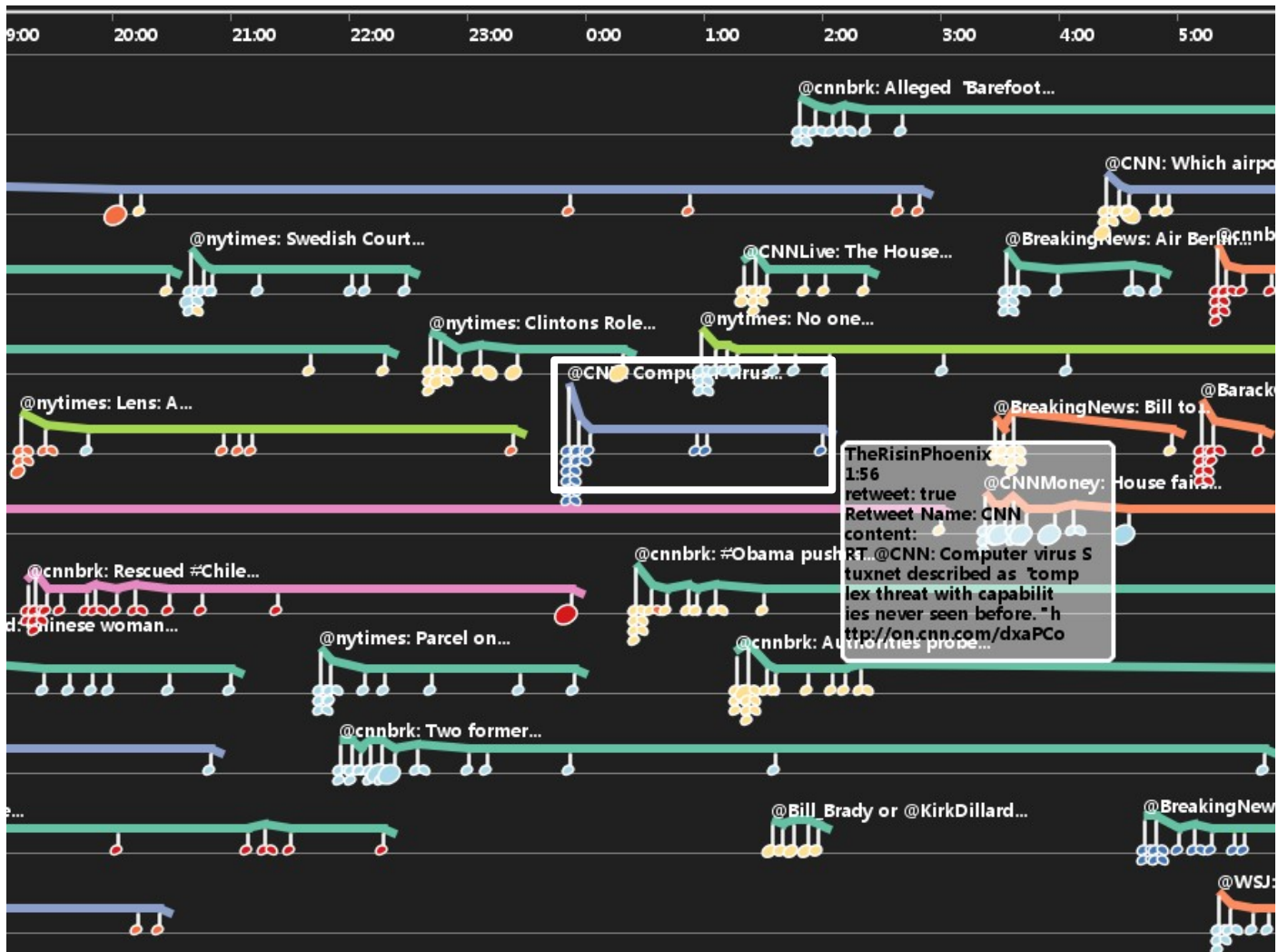


图 1 音符组表示 twitter 上的话题

2. 两种发布模式的可视化：转发，回复。

如图 2 所示，图中存在两种音符出现的模式，一种是话题刚刚出现时比较密集，随时间推移渐渐平缓的模式，如第六道的粉色音符组。

另外一种模式是话题出现后一直起伏不定的模式。如第四道粉色的一组音符。

转发

通过点击交互可以发现，第一种情况是转发模式，一般初始的这条消息来自与比较有影响力的人或机构，如图 2 中的 CNN 和 nytimes。而用户往往会在第一时间转发此条 twitter，越往后转发的次数越少，符尾越来越低。

转发的 twitter 可以继续被转发，这里使用点的大小来表示本微博被转发了多少次。

由于数据有限（原始数据集只包含 1600 多个用户的 twitter，而且暂时只提取的一天的数据），所以并不能发现多次转发引发的多轮话题的热潮。

至于转发广告扩大影响力的水军的情况，没在数据中发现。可能这是仍然因为数据样本量不大的原因，而且数据集很少有某公司产品的微博，而这正式水军比较多的微博类型之一。也有可能是因为我们使用的这一天的数据恰好就没有类似的情况发生。

回复

第二种情况是两人之间回复。具有朋友关系的两人往往会互相@，讨论某个话题。这个时候他们的微博往往是在一段时间内均匀出现，频率略有起伏。比如第一道的音符（因为说的不是英语所以不知道他们在讲什么话题）。

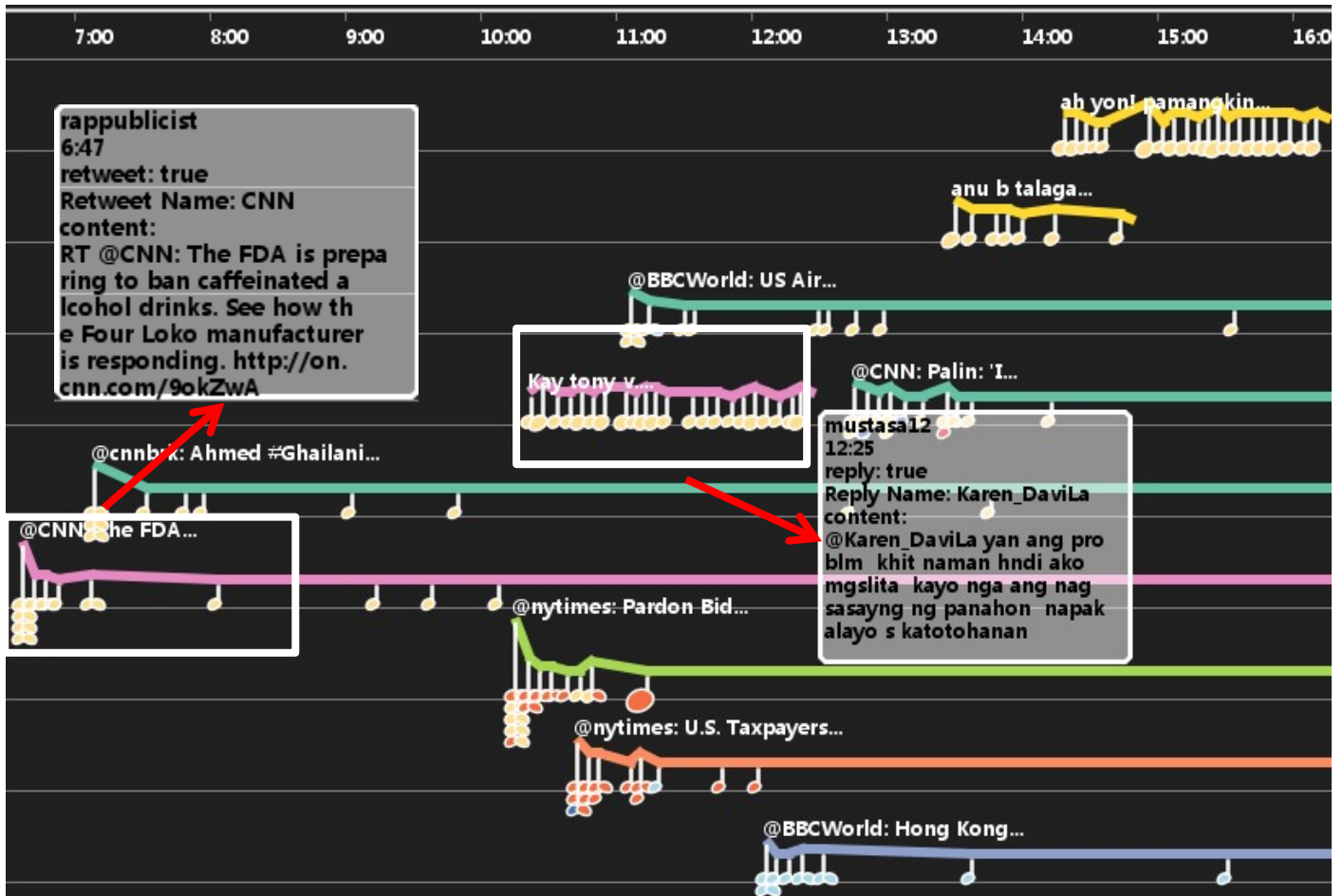


图 2 音符出现的模式表示了 twitter 上的两种发布模式

3. Twitter 用户的感情可视化。

为了显示，twitter 中用户的情感。使用红色系的音符表示 twitter 文本中表达的正面情绪，使用蓝色系的音符表示 twitter 用户表达的负面情绪，颜色越深表示其感情越强烈。而黄色表示中立的态度。

主要是从文本中分析出用户的情感，如果用户只是纯转发，则默认用户与新闻本身有相同的情感。这一部分的实现基本也是用过人工分类完成。

可以看到图 3 中第四道的音符组表示的政治类话题如下：“Palin: ‘I am’ thinking about 2012 bid.” 不同的用户表达量对这一事件的看法红色的点表示了强烈支持，蓝色的点表示了不满，而大部分人只是转发，可以默认为其持中立态度。

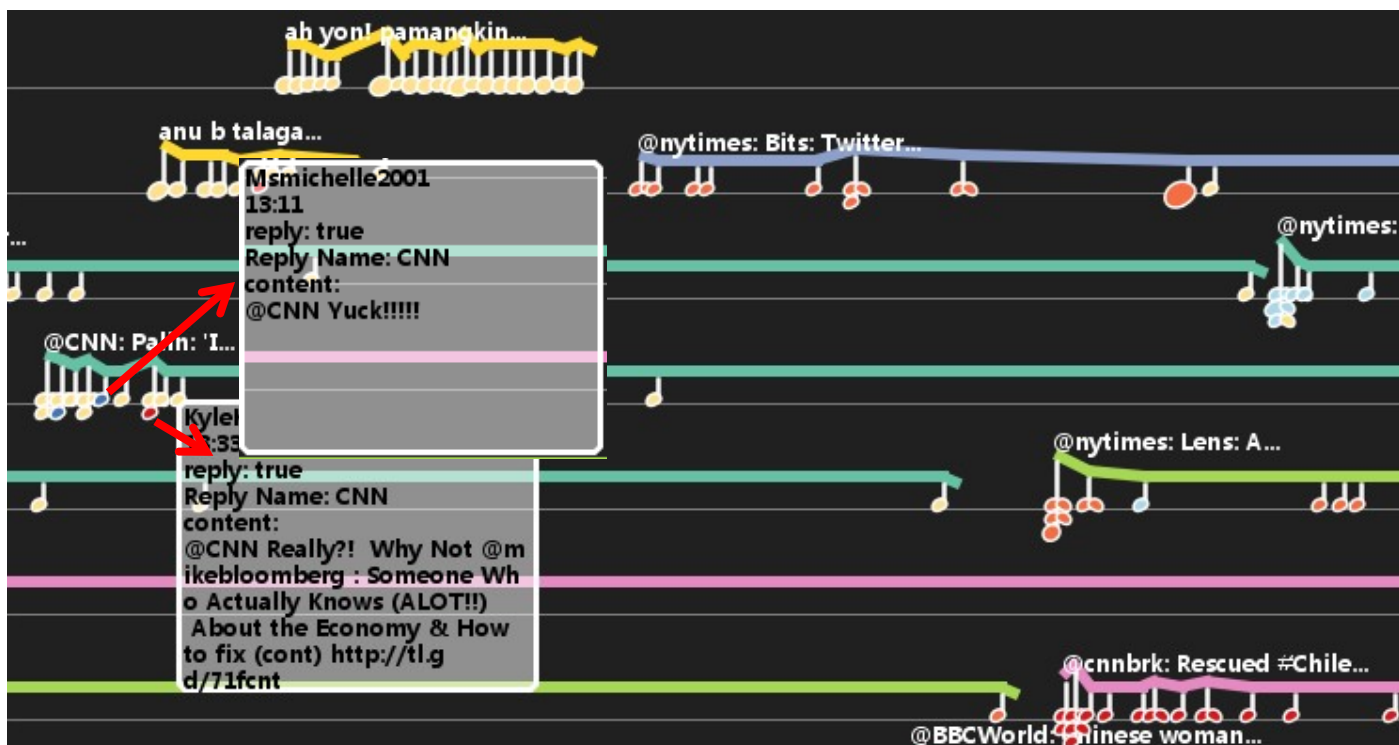


图 3 不同颜色的音符表示了不同人对该 tweet 的不同情感

总结

整体感觉效果要好于通讯数据的可视化。但是可以看出的时序模式不如淘宝交易数据种类多。从时序数据的角度来讲，基本就只有转发模式，回复模式和单独发模式。

Twitter 数据与其他时序用户行为数据相比最大的特点在于他的很多信息隐藏在文本数据中。因此，更多有意义的文本信息有待发掘并可视化。

Twitter 数据可视化现有问题：

1 纵向布局算法。

现在的纵向布局没有特定的含义，仅仅是让音符不重叠而已。

解决方案：

按话题相似性布局，话题越相似，音符组靠得越近。可以采用类似于聚类算法的方法实现。

按重要性布局，将重要的话题放在中间，按重要性向两边扩散。并且有必要实时调整各个话题的重要性，可能导致同一串音符不在同一条线上。可视化效果可能有点类似于 storyline。

2 文本挖掘算法。

现在的主题抽取，关键词分析，情感分析都是通过手动完成。

解决方案：

利用现有的文本算法解决。

3.数据不完整。

原始数据只提取了部分用户，很多的特定模式无法从数据中捕捉到。比如一些可能存在的情况，比如多次转发引发热潮，水军的机械式转发。

下周工作：

1. 组会论文报告
2. 改进 twitter 可视化中的文本分析方法。
3. 时序用户行为可视化英文文档书写